

An Automatic System for Acoustic Microphone Geometry Calibration based on Minimal Solvers

Simayijiang Zhayida, Simon Segerblom Rex, Yubin Kuang,
Fredrik Andersson, and Kalle Åström*
Centre for Mathematical Sciences, Lund University,
Box118, SE-22100, LUND, SWEDEN

October 10, 2016

Abstract

In this paper, robust detection, tracking and geometry estimation methods are developed and combined into a system for estimating time-difference estimates, microphone localization and sound source movement. No assumptions on the 3D locations of the microphones and sound sources are made. The system is capable of tracking continuously moving sound sources in an reverberant environment. The multi-path components are explicitly tracked and used in the geometry estimation parts. The system is based on matching between pairs of channels using GCC-PHAT. Instead of taking a single maximum at each time instant from each such pair, we select the four strongest local maxima. This produce a set of hypothesis to work with in the subsequent steps, where consistency constraints between the channels and time-continuity constraints are exploited. In the paper it demonstrated how such detections can be used to estimate microphone positions, sound source movement and room geometry. The methods are tested and verified using real data from several reverberant environments. The evaluation demonstrated accuracy in the order of few millimeters.

1 Introduction

Many of our everyday tools, such as smartphones, laptops, and tablet pc's, are equipped with microphones. Given the location of each one of these microphones, it is possible to use them as ad-hoc acoustic sensor network. Such sensor networks can be used for many interesting applications. One application is to improve the sound quality using so called beam-forming. Another application is so called speaker diarization, i.e. to determine who spoke when. Networks can also be used for localization and mapping.

*The work was supported by the strategic research projects ELLIT and eSENCE and the Swedish Foundation for Strategic Research (grant no. RIT15-0038), the Swedish Foundation for International Cooperation in Research and Higher Education, the Craaford Foundation and the Swedish Research Council.

If the microphone positions are unknown or only known to a certain accuracy, the results are inferior as is shown in [3].

Since the problem of acoustic microphone geometry calibration is essential to applications such as speaker localization, speaker tracking, indoor navigation, beamforming and speaker diarization, the problem has recently become a very active field of research, see [3], which gives an overview of the research field and their applications.

It is sometimes useful to think of the problem as having a signal processing part and a geometry estimation part. In the signal processing part one tries to obtain measurements from the raw audio measurements. These measurements will have both small measurement errors, but also gross errors or outliers. In the geometry estimation part one tries to estimate microphone and sound source positions from the noisy measurements extracted from the audio signals. There is also a correspondence problem. One has to track these sound events over channels and over time. This correspondence problem could be solved in the signal processing part, in the geometry estimation part or a combination of the two.

If the microphone positions are not known we obtain a chicken-and-egg kind of problem. Sound source localization is relatively easy if the microphone positions are known, and similarly microphone positions are relatively easy to solve if the sound source positions are known. But getting a sense of the data when the position of both microphones and sound sources are unknown is a more challenging problem. Furthermore if the sounds emitted are of unknown character and if the environment is reverberant, it is difficult to match the sounds from one sound source to the other.

The signal processing part is substantially easier if the sounds emitted are short in duration and if they have been specifically designed to be easily detected. Popular examples here are chirps or claps, cf. [1].

In this paper we present new methods for simultaneous signal processing and geometry estimation such as to obtain estimates of microphone positions, sound source positions and major reverberant planar structures, without requiring that the sound sources has a particular structure, for instance that they are of short duration.

Literature survey The literature within sound source localization using known microphone positions is very large, see e.g. [4–7], and the references therein.

If the sounds were generated by a speaker and if the speakers and the microphones are connected to the same sound card it would be possible to measure the Time-of-Arrival (TOA) directly, e.g. by using generalized cross correlation techniques between the emitted and received sound signals, [8]. Examples of such experiments are given e.g. in [9]. Similarly there are applications, e.g. using UWB or so called fine time measurements, also using a radio technique, which allows for time of arrival measurements, [10]. In these applications the signal processing part is implemented in hardware and TOA measurements are produced. For such applications, only the geometric estimation part needs to be solved.

Assuming that Time-of-Arrival (TOA) measurements are given, the problem of determining both microphone and sender positions is still challenging. Early attempts at solving this chicken-and-egg type of problem often resorted to adding extra assumptions on the location of some of the sensors or assumptions on one of the sound sources

being at the same position as one of the microphones, cf. [11–16].

Once a reasonable initial estimate of the positions are known, the problem can be formulated as a robust non-linear least squares problem, for which iterative methods exist, cf. [17, 18]. However, such methods depend on initialization and can get stuck in local minima. For a general graph structure, one can relax the TOA-based calibration problem as a semi-definite program [19].

For TOA measurements there is also considerable knowledge on minimal solvers. Minimal solvers for linearly constrained motion is given in [20]. The problem is well understood for the case of receivers or senders spanning sub-spaces of different dimension, [9]. When both receivers and senders span the same linear space, the minimal problems are 3 receivers and 3 senders in a plane, [21]. For the three-dimensional case the minimal problems are either 4 receivers and 6 senders (or symmetrically 6 receivers and 4 senders). The case of 5 receivers and 5 senders is close to minimal. Efficient algorithms for solving these problems are given in [22]. In [23] and refined in [20] a far field approximation was utilized to initialize both TOA and TDOA problems. In the far-field approximation, it is assumed that the distances between the speakers and receivers is considerable larger than between receivers.

Initialization of TDOA networks is studied in [24], where a solution to non-minimal case of 9 receivers and 4 speakers in 3D was presented.

In [25] a system is presented for calibrating two microphone arrays relative to each other. Here it is assumed that each array has been pre-calibrated first. The authors also assume that the environment is non-reverberant. This makes the signal processing easier. The authors calculate 3D positions of sound sources by maximizing the steered response power. This is essentially a 3D search, which is made by searching 3D space using a regular grid. Once a number of 3D matches are obtained the transformation between the two calibrated microphone arrays can be made.

In [26] it is assumed that there are several subgroups of pre-calibrated linear arrays, that all arrays and all sound sources are in a plane. Using beamforming one can determine the angle to the sound sources from each linear array. This gives rise to a 2D structure from motion problem, similar to that of 1D retinal computer vision, [27–29]. The authors propose a method of several random initializations of the unknown parameters followed by non-linear least squares optimization.

In [30], the authors introduce the idea of using the diffuse noise field for estimating the distance between two microphones is introduced. Given a diffuse noise field the coherence function is real valued and sinc-shaped with a parameter that depends on the distance between the two microphones. For this special case it is possible to estimate the parameter for every microphone pair and then use multi-dimensional scaling, [31], to obtain the microphone geometry. This was later used in [32] together with Random Sample Consensus (RANSAC) to obtain a more robust method.

This contribution In this paper, which is a combination and extension of two previous conference publications [1, 2], we focus on a system approach. For the system we assume that the microphones can be in unknown general 3D positions and that the microphones are synchronized. We do not put any constraints on the sound sources. They can also be in general 3D positions and do not have to fulfill the far-field crite-

ria. There can be multiple possible moving sound sources. The input to the system is a sound recording with m channels (one from each microphone). These are first processed to find the time-difference vectors. The estimated time-difference vectors typically contain noise, missing data and possible outliers. We then follow a stratified approach, where we first estimate offsets using a robust method. This is followed by a robust method for finding the 3D positions of all the microphones and the 3D positions of all sound sources. Next is refinement of matching matrix by finding more inliers from raw matches and also refinement of correlation pattern using sub-pixel methods, cf. [33].

The main contributions of this paper are:

- New efficient algorithms for solving for offsets given tdoa data for 5 minimal problems.
- New robust algorithms for solving the geometric TDOA problem given noisy data with outliers and missing data.
- New matching and tracking algorithms for obtaining TDOA data from acoustic signals.
- New robust algorithms for determining reverberant planes.
- Combination of these parts into a system.
- Development of a database of test cases with ground truth to test the system.

2 System Design

In this paper we present new tools and a first version of a system that can estimate (i) microphone positions, (ii) position and/or motion of one or several sound sources, and (iii) room geometry. We here assume that the number of simultaneous sound sources are few. In the experiments, we have mostly used a single moving sound source. We assume that the sound sources have relatively small spatial size and that the sounds are relatively omni-directional. We also assume that the microphones are relatively omni-directional. Furthermore we assume that major reverberant surfaces are few and planar, modelled as planes π_I . The system is tested in a real environment, e.g. as illustrated in Figure 1. For this reason, it is important that the system is robust to additional reverberations caused by other objects (chairs, people, etc) and robust to background sound sources, e.g. shoes and muffled talk.

Contrary to many previous papers, we do not assume the simpler case of microphones being placed close to each other, so that a far-field approximation can be used. Also we do not assume that the microphones are placed close to each other so that one may assume that they are within a wavelength of each other.

The input to the system consists of M sound recordings,

$$y_i(\tau T_s), \tau = 1, \dots, T, i = 1, \dots, M,$$

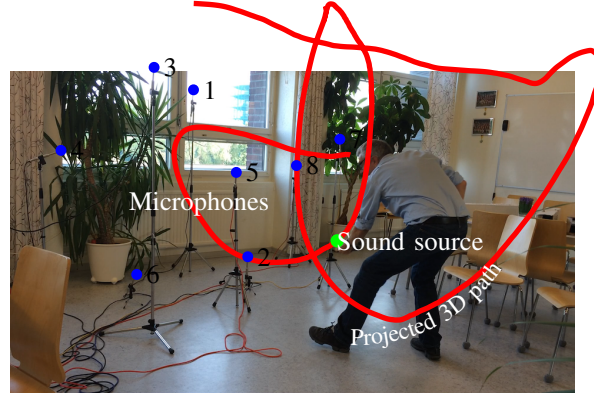


Figure 1: The paper presents a complete system for analysing sound recordings from m microphones. The figure illustrates one of the real experiments. In this experiment we used one moving sound source, shown as a red path. In the experiment there are reverberations in the floor and the window, and numerous weaker reverberations caused by chairs, tables, humans etc. The system calculates microphone positions, room geometry and sound source motion.

where i is used as index for the recordings and τ denotes the index for the sampling point and T_s is the sampling period. The sound sources are assumed to be spatially small and omnidirectional, so that they can be modelled as points moving in space. Each such sound event have an onset time t_0 and offset time t_1 . During this interval $t \in [t_0, t_1]$ the sound source moves in space according to $\mathbf{s} : [t_0, t_1] \rightarrow \mathbb{R}^3$. during this interval the emitted sound is given by $x(t)$. According to the model, at each time, the measured signals

$$y_i(t) \approx \sum_k \alpha_{i,k}(t) x(t - \tau_{i,k}(t)), \quad (1)$$

is approximately a sum of a few scaled and time-delayed versions of the sound $x(t)$, emitted by the sound source plus noise, e.g. by stray sounds and reverberations not modelled by the main reflective planes. The time-delays $\tau_{i,k}$ are themselves time dependent. However, during the analysis of short time-intervals, they can be considered to be almost constant. Here it is assumed that the speed of the sound sources is much less than the speed of sound. We use an index k to denote the index of the reflections, where $k = 1$ is used to denote the direct path and higher indices are used for single and then multiple reflections in the planar surfaces.

Let \mathbf{r}_i , $i = 1, \dots, m$ be the spatial coordinates of the m receivers. For the case of acoustic mirrored microphone we will use the notation $\mathbf{r}_{i,k}$ for the mirrored version of microphone i corresponding to multi-path component k . Let \mathbf{s} , be the spatial coordinate vector of a position at which a sound was emitted for a sound event at a particular time. For measured time of arrival τ_{ik} from sound transmission at a position \mathbf{s} with unknown emission time t_{emission} to receiver at a position $\mathbf{r}_{i,k}$, we have that $v\tau_{i,k} = \|\mathbf{r}_{i,k} - \mathbf{s}\|_2$, where v is the speed of measured signals and $\|\cdot\|_2$ is the l^2 -norm. The speed v is assumed to be known and constant.

If the sounds were generated by a speaker and if the speakers and the microphones are connected to the same sound card it would be possible to measure these time of arrivals directly. This is possible since then both functions y and x in Equation 1 would be known. So by generalized cross correlation techniques it is possible to determine $\tau_{i,k}$ from x and y , [8]. In this paper, however, we consider the problem of listening to ambient sound. We assume that the microphones (the receivers) are connected to the same sound card, or that they are synchronized. This assumption could in fact be relaxed using calibration techniques to synchronize the data afterwards, e.g. using techniques developed in [34]. This extension would allow the use of arbitrary non-synchronized recording devices. For the rest of the paper, however, we assume that the receivers are synchronized but that the sound sources are unknown and thus unsynchronized. Thus, we are forced to use time difference measurements.

Two channels can be matched using different techniques, e.g. cross-correlation or generalized cross-correlation. Generalized cross-correlation method can weaken the impact of noise on the delay estimation accuracy [8, 35]. Here we opted to use GCC-PHAT for the estimation of time-differences. The matching score between sounds i_1 and i_2 typically have peaks for relative time delays $\tau_{i_2,k_2} - \tau_{i_1,k_1}$, for different combinations of multi-path components k_1 and k_2 for the two microphones in question, see (Equation 1). Our experience is that the difference for the direct path, i.e. $\tau_{i_2,1} - \tau_{i_1,1}$ is more likely have a high matching score. However, due to noise it is often the case that this direct path does not yield the strongest peak and sometimes the direct path does not yield a peak at all. In this paper we will usually scale the time-differences with speed so as to obtain range-differences. The connection between the range difference and the geometric positions of the sound source and the receivers are then

$$w_{i_1,k_1,i_2,k_2} = v(\tau_{i_2,k_2} - \tau_{i_1,k_1}) = \|\mathbf{s} - \mathbf{r}_{i_2,k_2}\| - \|\mathbf{s} - \mathbf{r}_{i_1,k_1}\|. \quad (2)$$

In principle one could study correlation techniques that study three or more channels directly. But the search space and computational complexity then becomes a problem. For two channels each time-difference has one degree of freedom. This one degree of freedom is determined efficiently using the fast Fourier transform. For m channels each multi-path component has $m - 1$ degrees of freedom. Our hypothesis is that using pairwise channels gives the best trade-off between computational complexity and performance. However, in the future, it would be interesting to further investigate techniques that explicitly study multi-channel correlation techniques.

In the sequel we will use the *matching vector* for time matchings of the same signal at some time instant in each channels, which is denoted as $(u_1, u_2, \dots, u_m)^T$. We will allow missing data for such vectors, i.e. there might be one or several indices in a vector that has unknown values. The components of the vector might also contain outliers. Also introduce $o = c(t_1 - t_0) = \|\mathbf{r}_1 - \mathbf{s}\|$ as the offset. This can be interpreted as the distance from the sound to microphone 1. Using this notation the measurement equation (1) becomes

$$u_i = \|\mathbf{r}_i - \mathbf{s}\| - o.$$

Let j be used as an index for different sounds. The key idea is that using a number of such measurements u_{ij} it is possible to estimate the unknown parameters $(\mathbf{r}_i, \mathbf{s}_j, o_j)$ so

that

$$u_{i,j} = ||\mathbf{r}_i - \mathbf{s}_j|| - o_j.$$

3 Data collection

We have made several experiments with 8 microphones. The measured data (i.e. received signals at the microphone array) was obtained by 8 microphones (Shure SV100 or T-bone MM-1) which are connected to an audio interface (M-Audio Fast Track Ultra 8R), connected to a laptop. The microphones were positioned in a room with approximate distance 0 - 3 meters from each other. The input to the system is a sound recording with M channels (one from each microphone).

The M sound channels were sampled at 96000 Hz. We assume speed of sound is approximately $c = 343m/s$ for room temperature and normal atmospheric pressure.

Sounds can be generated in several scenarios, for e.g.:

- Random distinct sound bursts made for example by banging two spoons together. This produces a set of discrete sound events that are relatively easy to detect and match.
- One continuously moving sound source playing part of a song. This produces a set of smoothly changing time-differences.
- Several continuously moving sound sources.
- Mixture of several people talking, clapping, walking around in the room. The sound sources appear and disappear, humans start and stop talking.
- In anechoic chamber
- Lunch room.
- Lecture room.



(a) Microphones.



(b) Computer and audio interface.

Figure 2: Hardwares for data collection.

For the datasets, we have also obtained ground truth microphone positions and sound source motion.

4 Time-Difference Estimation and Tracking

Experiments were made in different environments (normal and echo-free). Different types of sound sources were used (claps, voices, continuous songs). Continuously moving sound sources are substantially more challenging.

For both signal processing components the output is an estimate of range difference w_{i_1, k_1, i_2, k_2} for the direct path $(k_1, k_2) = (1, 1)$ at a number of time instants $t_j, j = 1, \dots, n$.

4.1 Time-Difference Estimation - Claps

For claps, we use flank detectors that detect the onset times t_i of such claps for each channel i . We then use a simple matching scheme to match the onset times along the different sound channels. Here we assume that the different claps do not occur too frequently.

Assume that there are n claps, and that each clap is index by $1 \leq j \leq n$. For each such clap we thus get m onset times t_1, \dots, t_m for the m channels. Assuming that the onset times correspond to the arrival of the direct component of the sound wave we thus get range difference estimates

$$w_{1,1,i,1} = v(t_i - t_1). \quad (3)$$

Introduce the column vector $\mathbf{u} \in \mathbb{R}^m$, whose components are

$$u_i = w_{1,1,i,1} = v(t_i - t_1). \quad (4)$$

Notice that by definition $u_1 = 0$. Each sound clap j gives rise to such a different matching vector \mathbf{u}_j . Now define the $m \times n$ matrix

$$\mathbf{U} = (\mathbf{u}_1 \quad \dots \quad \mathbf{u}_n), \quad (5)$$

whose columns correspond to the matching vectors of the n claps. Each sound clap j has a 3D position vector \mathbf{s}_j . Notice that

$$\mathbf{U}_{ij} = \|\mathbf{r}_{i,1} - \mathbf{s}_j\| + o_j, \quad (6)$$

with $o_j = -\|\mathbf{r}_{1,1} - \mathbf{s}_j\|$. The estimation of o_j , \mathbf{s} and \mathbf{r} are discussed in Sections 5 and 6.

4.2 Time-Difference Estimation - Continuously moving sound sources

Each sound file is divided into frames, 2048 sample points long, 1000 sample points apart. The Generalized Cross Correlation with Phase Transform (GCC-PHAT) is calculated between corresponding frames for all microphone pairs. Arranging the correlation score of each frame as a column in a matrix yields a score matrix as illustrated in Figure 3. In the figure the x-axis represents time into the recording in seconds and the y-axis are range-differences $v(\tau_{m_2} - \tau_{m_1})$ in meters. As can be seen in the figure the data is quite noisy. Nevertheless, certain patterns in the data

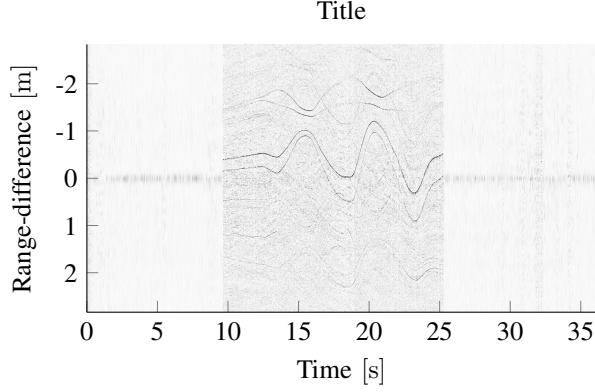


Figure 3: GCC-PHAT between all frames for two channels (channel 1 vs channel 2). Each pixel in column represents one correlation value of frames from channel 1 and 2. Only positive values plotted. The most prominent curve structure is the result for the direct sound path.

can be discerned. These curves correspond to direct and indirect range-differences $w_{i_1, k_1, i_2, k_2} = \|\mathbf{s} - \mathbf{r}_{i_2, k_2}\| - \|\mathbf{s} - \mathbf{r}_{i_1, k_1}\|$, for different combinations of k_1 and k_2 , see Equation 2.

In general, the received signal model of two microphones can be expressed as

$$f_1(t) = g(t) + n_1(t),$$

$$f_2(t) = g(t - \tau) + n_2(t),$$

where $n_1(t)$ and $n_2(t)$ are noise and τ is the time difference of the sound source to the microphone pair. The generalized cross-correlation or GCC function between $f_1(t)$ and $f_2(t)$ is

$$R_{y_1 y_2}(\tau) = \int_{-\infty}^{\infty} W(\omega) G_{f_1 f_2}(\omega) e^{j2\pi\omega\tau} d\omega,$$

where $G_{f_1 f_2}$ is the cross-power spectrum of received signals of two microphones, and the generalized frequency weighting is

$$W(\omega) = H_1(\omega) H_2^*(\omega) = \frac{1}{|G_{f_1 f_2}(\omega)|},$$

where $*$ denotes the complex conjugate.

The main idea here is that some of the peaks in Figure 3 correspond to range-difference measurements

$$w_{i_1, k_1, i_2, k_2} = \|\mathbf{s} - \mathbf{r}_{i_2, k_2}\| - \|\mathbf{s} - \mathbf{r}_{i_1, k_1}\|. \quad (7)$$

The difficulty lies in determining which peaks are relevant and for each such peak to determine the corresponding multi-path index pair (k_1, k_2) . If such correspondences

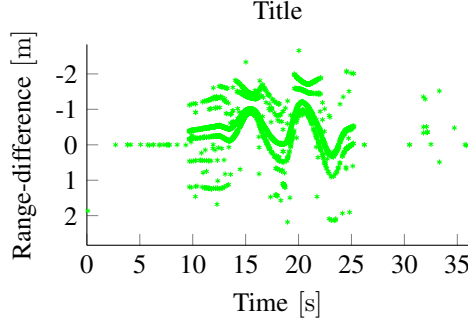


Figure 4: Peaks after matching using data from all channels.

can be found the next step is the geometric estimation of microphone and sound source positions.

The strategy used here is to use heuristics to try to first find peaks that correspond to the direct path components, i.e. those with multi-path index pair $(k_1, k_2) = (1, 1)$.

4.3 Peak selection

The four strongest peaks for each frame from the cross-correlation are selected. Only peaks above a certain threshold are considered to reduce the number of outliers. Many of these peaks correspond to direct path range differences $w_{i_1,1,i_2,1}$. For a continuously moving sound source these peaks typically generate smooth tracks in the figure. As can be seen in the figure, the set of detections contain a mixture of true positives and false positives. The false positives consists of both noise but also of range difference for multi-path components, i.e. range differences w_{i_1,k_1,i_2,k_2} with multi-path index pair $(k_1, k_2) \neq (1, 1)$. The number of false positives are evaluated in the experimental evaluation, (Section 9.4).

4.4 Peak tracking

Assuming that the motion of the sound source is relatively smooth (at least at times), we exploit the continuity constraint by grouping together range difference peaks. Here we use a non-repeating RANSAC algorithm to find so called *tracklets*, i.e. small track parts of the curves. These are 21 frames wide with 1 frame overlap. Peaks closer than a threshold from the line are considered to be inliers. Both the lines with the most inliers and the inliers themselves are stored for each tracklet. If two lines share more than one inlier, the one with less inliers is ignored. The result is illustrated in Figure 5. Notice that the process further removes outlier and errors while keeping the dominant range difference tracks in the data.

In the next step, all line segments closer than a certain threshold are connected to form longer tracks. Segments separated by a small amount in time sharing roughly the same line equation are connected to form even longer segments. The hypothesis is that the longest segment corresponds to the direct path.

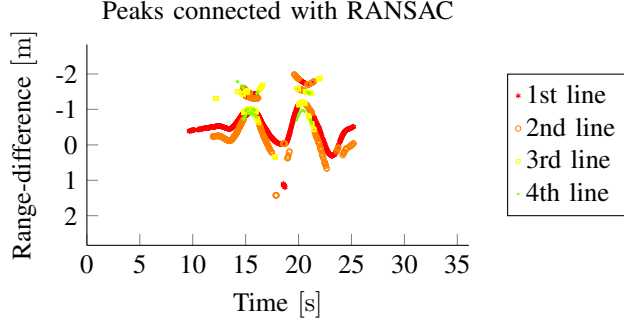


Figure 5: Inliers for line segments found with RANSAC.

Thus we obtain an hypothesis of the range difference for the direct path $w_{i_1,1,i_2,1}$ at a number of instances $t_j, j = 1, \dots, n$ along the time axis. For each time instance t_j we collect the range difference estimates $w_{1,1,i,1}$ in a column vector, $\mathbf{u} \in \mathbb{R}^m$, i.e. so that the components are given by

$$u_i = w_{1,1,i,1}. \quad (8)$$

Now define the $m \times n$ matrix

$$\mathbf{U} = (\mathbf{u}_1 \quad \dots \quad \mathbf{u}_n), \quad (9)$$

whose columns correspond to the matching vectors of the n time instants t_j . Again notice that

$$\mathbf{U}_{ij} = \|\mathbf{s}_j - \mathbf{r}_{i,1}\| + o_j, \quad (10)$$

with $o_j = -\|\mathbf{s}_j - \mathbf{r}_{1,1}\|$. The estimation of o_j , \mathbf{s} and \mathbf{r} are discussed in Sections 5 and 6.

5 Offset Estimation

From the signal processing step we obtain a $m \times n$ matrix \mathbf{U} of range difference estimates \mathbf{U}_{ij} . These are linked to the positions of the m microphones $\mathbf{r}_{1,1}, \dots, \mathbf{r}_{m,1}$, the n sound events $\mathbf{s}_1, \dots, \mathbf{s}_n$ and the n offsets o_1, \dots, o_n according to

$$\mathbf{U}_{ij} = \|\mathbf{s}_j - \mathbf{r}_{i,1}\| + o_j, \quad (11)$$

where the offsets are defined as

$$o_j = -\|\mathbf{s}_j - \mathbf{r}_1\|. \quad (12)$$

In order not to clutter the notation in this section we are going to drop the multi-path index, i.e. write $\mathbf{r}_i = \mathbf{r}_{i,1}$.

In this section we derive the rank constraint, which can be used for estimating the offsets o_j from the elements of matching vectors \mathbf{U}_{ij} . This is an extension of the work in [1, 36].

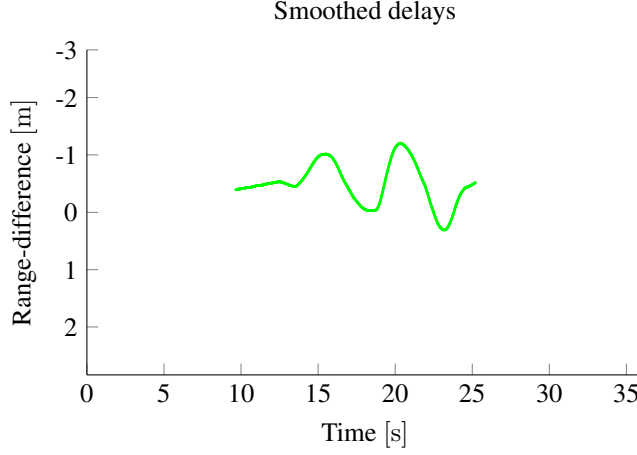


Figure 6: Peaks after smoothing.

5.1 The Rank Constraint

To begin with, let us rewrite the measurement equation

$$\mathbf{U}_{ij} = \|\mathbf{s}_j - \mathbf{r}_i\| + o_j. \quad (13)$$

as

$$(\mathbf{U}_{ij} - o_j)^2 = \|\mathbf{r}_i - \mathbf{s}_j\|^2 = \mathbf{r}_i^T \mathbf{r}_i - 2\mathbf{r}_i^T \mathbf{s}_j + \mathbf{s}_j^T \mathbf{s}_j.$$

A simple rearrangement yields

$$\mathbf{U}_{ij}^2 - 2\mathbf{U}_{ij}o_j^2 = \mathbf{r}_i^T \mathbf{r}_i - 2\mathbf{r}_i^T \mathbf{s}_j + (\mathbf{s}_j^T \mathbf{s}_j - o_j^2).$$

By constructing the vectors $\tilde{\mathbf{R}}_i = [1 \quad \mathbf{r}_i^T \quad \mathbf{r}_i^T \mathbf{r}_i]^T$ and $\tilde{\mathbf{S}}_j = [\mathbf{s}_j^T \mathbf{s}_j - o_j^2 \quad -2\mathbf{s}_j^T \quad 1]^T$, we obtain

$$\mathbf{U}_{ij}^2 - 2\mathbf{U}_{ij}o_j^2 = \tilde{\mathbf{R}}_i^T \tilde{\mathbf{S}}_j. \quad (14)$$

By collecting $\tilde{\mathbf{R}}_i$ and $\tilde{\mathbf{S}}_j$ into the matrices $\tilde{\mathbf{R}} \in \mathbb{R}^{5 \times m}$ and $\tilde{\mathbf{S}} \in \mathbb{R}^{5 \times n}$, we have $\mathbf{D} = \tilde{\mathbf{R}}^T \tilde{\mathbf{S}}$, where \mathbf{D} is the $m \times n$ matrix with elements $d_{ij} = u_{ij}^2 - 2u_{ij}o_j$. This gives a matrix \mathbf{D} that is at most of rank 5 as we increase m and n .

Now form the matrix $\mathbf{F} = \frac{1}{-2} \mathbf{C}_M^T \mathbf{D} \mathbf{C}_N$, where $\mathbf{C}_m = [-\mathbf{1}_{m-1} \quad \mathbf{I}_{m-1}]^T$, $\mathbf{C}_n = [-\mathbf{1}_{n-1} \quad \mathbf{I}_{n-1}]^T$, and where $\mathbf{1}_{n-1}$ is a $(n-1) \times 1$ vector with 1 as entries and \mathbf{I}_{n-1} is identity matrix of size $(n-1)$. This matrix \mathbf{F} has size $(m-1) \times (n-1)$. The elements are $f_{i,j} = \frac{1}{-2}(\tilde{d}_{i,j} - \tilde{d}_{i,1} - \tilde{d}_{1,j} + d_{1,1})$ as shown below.

$$\mathbf{F} = \frac{1}{-2} \begin{pmatrix} -1 & 1 & & & \\ -1 & & 1 & & \\ \vdots & & & \ddots & \\ -1 & & & & 1 \end{pmatrix} \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{m1} & d_{m2} & \cdots & d_{mn} \end{pmatrix} \begin{pmatrix} -1 & -1 & \cdots & -1 \\ 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix}$$

$$\begin{aligned}
&= \frac{1}{-2} \begin{pmatrix} -d_{11} + d_{21} & -d_{12} + d_{22} & \cdots & -d_{1n} + d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -d_{11} + d_{m1} & -d_{12} + d_{m2} & \cdots & -d_{1n} + d_{mn} \end{pmatrix} \begin{pmatrix} -1 & -1 & \cdots & -1 \\ 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix} \\
&= \frac{1}{-2} \begin{pmatrix} d_{11} - d_{21} - d_{12} + d_{22} & \cdots & d_{11} - d_{21} - d_{1n} + d_{2n} \\ \vdots & \ddots & \vdots \\ d_{11} - d_{m1} - d_{12} + d_{m2} & \cdots & d_{11} - d_{m1} - d_{1n} + d_{mn} \end{pmatrix}
\end{aligned}$$

It is relatively easy to see that $\text{rank}(\mathbf{F}(o)) \leq 3$. In fact

$$\mathbf{F}(o) = \mathbf{R}^T \mathbf{S},$$

with

$$\mathbf{R} = [\mathbf{r}_2 - \mathbf{r}_1, \dots, \mathbf{r}_m - \mathbf{r}_1]$$

and

$$\mathbf{S} = [\mathbf{s}_2 - \mathbf{s}_1, \dots, \mathbf{s}_n - \mathbf{s}_1].$$

The matrix $\mathbf{F}(o)$, here called the *double compaction matrix*, depends on the matching vectors \mathbf{U}_j and on the offsets o_j . Notice here that if the dimension of the affine span of either the microphones or the the sound events are less than 3, then the rank will be lowered. For example if all the receivers lie in a plane, then $\text{rank } \mathbf{R} \leq 2$ and consequently $\text{rank } \mathbf{F} \leq 2$. Another example is if all the sound events are on a line then $\text{rank } \mathbf{S} \leq 1$ and consequently $\text{rank } \mathbf{F} \leq 1$.

5.2 Minimal solvers for the rank constraint

The problem of determining the offsets o now becomes that of finding o so that $\text{rank}(\mathbf{F}(o)) \leq 3$. The elements of \mathbf{F} are linear in o . The fact that the rank is less than 3 means that all 4×4 sub-determinants of \mathbf{F} are zero. These constraints on these sub-determinants or minors are thus polynomial constraints of degree 4 in o .

For a $(m-1) \times (n-1)$ matrix $\tilde{\mathbf{F}}$ of rank 3, the number of constraints $N_c = |\Lambda_4| = \binom{m-1}{4} \binom{n-1}{4}$ among which $(m-4)(n-4)$ constraints are linearly independent. Each constraint is a polynomial equation of degree 4 in $\{o_1, \dots, o_n\}$. For different choices of m and n , this system of polynomials equations can either be well-defined, over-determined or under-determined. To resolve this, we rely on algebraic geometry tools and make use of *Macaulay2* [37].

It turns out that there are several choices for m and n that produce well-defined and solvable polynomial systems. We summarize those cases and the number of solutions of the related polynomial systems for $K = 3$ and $K = 2$ in Table 1. In the following discussion, we denote the case with m receivers and n transmitters as m/n s. Given these solvable cases, we can apply numerically stable polynomial solvers based on methods described in [38] to solve for the unknown offsets. Using this technique, solving the system of polynomial equations is converted to an eigenvalue problem, which

K	m	n	N_{sol}
3	9	5	1
3	7	6	5
3	6	8	14
2	7	4	1
2	5	6	5

Table 1: Number of solutions to the polynomial systems of rank constraints on unknown offsets for different cases in (III) (3D and 2D).

is efficiently solved by techniques from numerical linear algebra. The resulting solvers are fast (execution time in the order of milliseconds) and are non-iterative (in the sense that they only use standard numerical linear algebra routines and no iterations).

5.3 Robust estimation of offsets

The resulting matching \mathbf{U} from the signal processing steps include (i) inlier measurements with noise, (ii) outliers due to gross errors in matching and (iii) missing data. Robust estimation of offsets can be achieved by algorithms that try to find the offsets with the most inliers (or the fewest outliers), for this propose we use RANSAC, cf. [39]. It has proved to be useful for many difficult parameter estimation problems in signal processing and computer vision, e.g. [10, 25, 26], since the methods are robust to both measurement errors and outliers. RANSAC select the solution with the most inliers and among these the one with the lowest error score. The algorithm requires two parameters, the agreement threshold (how close does an inlier have to be) and number of iterations.

The main idea of RANSAC is to use an hypothesize and test paradigm. First one hypothesize a minimal subset of data for which it is possible to solve for the parameters (the offsets in this case). Using this hypothesized solution one then extends the solution to the rest of the data and determines how many of the remaining data that agrees with the solution. Here it is important that both the minimal solvers and the verification step is fast, since this allows for more iterations. All of the minimal solvers that we have generated (Table 1) are potentially useful. Here we explain the methodology using the 7 microphone, 6 sound events (rank 3) solver.

1. From the set of matching vectors, randomly select a subset of 7 channels and 6 matching vectors, and solve for the offset for the minimal case 7×6 above. Use the closed form algorithm for finding the offsets o_j for these 6 matching vectors. The relevant (valid) solutions should have offsets that are real, and since $\|\mathbf{r}_i - \mathbf{s}_j\| = u_{ij} - o_j$, the offsets should fulfill the constraints $u_{ij} \geq o_j$ for $i = 1, \dots, M$ and $j = 1, \dots, N$. Ignore solutions that do not fulfill these constraints.
2. Extend to the remaining microphones, and solve for $M \times 6$.

3. For each solution study how many of the remaining matching vectors that fulfill the geometric constraint, keep largest set of inliers. The measurements treat as inliers if $|\mathbf{d}_{ij} - \|\mathbf{m}_i - \mathbf{s}_j\|_2| \leq \epsilon$ for a given ϵ .
4. Repeat (1) to (3) a fixed number of times and choose the solution with the maximum number of inlier matching vectors.

5.4 Non-linear optimization

Given an initial estimate of offsets it is possible to improve on the offsets using non-linear optimization. Typically this is applied once a rough initial estimation has been found, such initial estimates may be obtained in a number of different ways including use of minimal solvers in a RANSAC framework, estimation of parameters using factorization techniques, etc.

5.5 Rank-based Nonlinear Optimization

In this section, we derive a iterative nonlinear optimization method for improving the estimate of the unknown offsets o_j . While the non-iterative schemes presented above apply only to specific number of M and N with no missing data, the method present in this section copes with such cases naturally.

Given the knowledge that the measurement matrix after compaction is of rank K , we can derive another scheme based nonlinear optimization to estimate the offsets $\mathbf{o} = (o_1, o_2, \dots, o_N)$. The idea is to find the offset such that the measurement matrix after compaction is as close to a rank- K matrix as possible. Thus, we have the following minimization problem:

$$\begin{aligned} \min_{\mathbf{o}, \mathbf{A}} \quad & \|\mathbf{F}(\mathbf{U}, \mathbf{o}) - \mathbf{A}\|_{F, \Omega}, \\ \text{s.t.} \quad & \text{rank}(\mathbf{A}) = K, \end{aligned} \tag{15}$$

where \mathbf{F} is the matrix resulting from the compaction operators as in the previous section, \mathbf{U} is matching matrix, $\mathbf{A} \in \mathbb{R}^{(m-1) \times (n-1)}$, and $\|\cdot\|_{F, \Omega}$ is the Frobenious norm on the matrix entries that are observed specified by the set Ω .

An alternative formulation is to optimize over the unknown offsets and model the errors direction on the measurement matrix as follows.

$$\begin{aligned} \min_{\mathbf{o}, \tilde{\mathbf{U}}} \quad & \|\mathbf{U} - \tilde{\mathbf{U}}\|_{F, \Omega}, \\ \text{s.t.} \quad & \text{rank } \mathbf{F}(\tilde{\mathbf{U}}, \mathbf{o}) = K. \end{aligned} \tag{16}$$

Here $\tilde{\mathbf{U}}$ is approximation of matching matrix. Similar formulation of the minimization problem (15) has been proposed in [40] to utilize the rank constraints. Given that rank constraint on \mathbf{A} , the minimization problem is non-convex. In [40], an alternating scheme is proposed. To be more specific, one first fixes the offsets \mathbf{o} , and solve for the optimal \mathbf{A} using SVD. Then, one fixes \mathbf{A} , the problem of finding the optimal \mathbf{o} is convex. However, the rate of convergence of this alternative scheme is very slow. Thus, an

additional regularization term on \mathbf{A} is introduced to speed up the convergence. Here, we used a gradient descent scheme that utilize a local parameterization of the rank constraints on \mathbf{A} directly.

The optimization problem is basically a non-linear least squares problem. Here we adopt a strategy to make local parametrization at each iteration step. For the optimization problem (15), this is implemented as follows. Assuming that the estimate for the current iteration $(\mathbf{o}_k, \mathbf{A}_k)$. A local parametrization for the offsets is straightforward. Introduce a local perturbation (y_1, \dots, y_n) , where x_i denotes the change in offset number i . For the local parameterization of \mathbf{A}_k , we use singular value decomposition to rewrite \mathbf{A}_k as

$$\mathbf{A}_k = U E_K W,$$

where U is the $m \times m$ unitary matrix from the singular value decomposition $\mathbf{A}_k = U S V^T$, where W is the $K \times n$ matrix corresponding to the first K rows of $S V^T$ and where E_K is a $m \times K$ matrix, the first K columns of a $m \times m$ identity matrix. A local parametrization is then

$$\mathbf{A}_{k+1} = U e^{\sum_i^n z_i B_i} E_K \left(\sum_{j=1}^{Kn} (W + w_j C_j) \right).$$

Here C_j are a basis for matrices of size $K \times n$, where B_i are a basis for $m \times m$ anti-symmetric matrices with zeros in the upper-left $K \times K$ block and zeros in the lower-right $(m - K) \times (m - K)$ block, as shown below:

The parametrization involves $m + K(m - K) + Kn$ parameters

$$x = \begin{pmatrix} y \\ z \\ w \end{pmatrix}.$$

It is relatively straightforward to calculate the analytic derivatives of A_{k+1} with respect to x . And also the derivatives of $F(\mathbf{U}, \mathbf{o})$ with respect to \mathbf{o} . This makes it possible to calculate the residuals and the jacobian of the non-linear least squares problem and implement an efficient Gauss-Newton optimization scheme for solving (15).

We then use the minimal solver on random subsets of data. If the solution is consistent with many of the remaining data, we extend the solution and then refine it using the iterative methods. This makes it possible to handle both missing data and outliers.

6 Estimation of Microphone and Sound Source Positions

Once we have calibrated the measurement matrix with the offsets $\{o_j\}$, we can obtain distance measurements $d_{ij} = \mathbf{U}_{ij} - o_j$. The measurement equation then becomes

$$d_{ij} = \mathbf{U}_{ij} - o_j = \|\mathbf{s}_j - \mathbf{r}_{i,1}\|. \quad (17)$$

We then proceed to solve the locations of microphones $\{\mathbf{m}_i\}$ and sounds $\{\mathbf{s}_j\}$ as a TOA problem. Note that one can only reconstruct locations of microphones and sounds up to Euclidean transformation and mirroring. The TOA-based self-calibration problem studied here corresponding to bipartite graph.

Similar to before, by squaring the measurement equations,

$$d_{ij}^2 = (\mathbf{r}_i - \mathbf{s}_j)^T (\mathbf{r}_i - \mathbf{s}_j) = \mathbf{r}_i^T \mathbf{r}_i + \mathbf{s}_j^T \mathbf{s}_j - 2\mathbf{r}_i^T \mathbf{s}_j \quad (18)$$

we obtain a set of mn polynomial equations in the unknown. We form mn new equations by linear combinations of the ones above. In particular we choose equations according to:

$$d_{11}^2 = (\mathbf{r}_1 - \mathbf{s}_1)^T (\mathbf{r}_1 - \mathbf{s}_1), \quad (19)$$

$$d_{1j}^2 - d_{11}^2 = -2\mathbf{r}_1^T (\mathbf{s}_j - \mathbf{s}_1) + \mathbf{s}_j^T \mathbf{s}_j - \mathbf{s}_1^T \mathbf{s}_1 \quad (20)$$

$$d_{i1}^2 - d_{11}^2 = -2(\mathbf{r}_i - \mathbf{r}_1)^T \mathbf{s}_1 + \mathbf{r}_i^T \mathbf{r}_i - \mathbf{r}_1^T \mathbf{r}_1, \quad (21)$$

$$\frac{d_{ij}^2 - d_{i1}^2 - d_{1j}^2 + d_{11}^2}{-2} = (\mathbf{r}_i - \mathbf{r}_1)^T (\mathbf{s}_j - \mathbf{s}_1). \quad (22)$$

By this process there are $(m-1)(n-1)$ equations of type (22) which are bilinear in \mathbf{r} and \mathbf{s} , which has rank 3 as discussed before.

By factorizing $\hat{\mathbf{B}} = \tilde{\mathbf{R}}^T \tilde{\mathbf{S}}$, we can almost solve the node calibration problem, however there are a few remaining unknowns:

- Since the factorization is not unique. If $\hat{\mathbf{B}} = \tilde{\mathbf{R}}^T \tilde{\mathbf{S}}$, then also $\hat{\mathbf{B}} = \tilde{\mathbf{R}}^T A A^{-1} \tilde{\mathbf{S}}$ is a valid factorization.
- The starting sender positions \mathbf{s}_1 is unknown.
- The starting receiver positions \mathbf{r}_1 is unknown.

However, since the choice of coordinate system is arbitrary anyway, one may without loss of generality set the origin to \mathbf{r}_1 . Also since any matrix A can be QR-factorized as a rotation matrix times a triangular matrix, one may assume that A is triangular, i.e. $A = L$. Thus fixing the rotation of the coordinate system. The remaining unknowns can be parametrized as

$$\begin{aligned} \mathbf{r}_1 &= \mathbf{0}, \\ \mathbf{s}_1 &= \mathbf{L}\mathbf{b}, \\ \mathbf{r}_i &= \mathbf{L}^{-T} \tilde{\mathbf{R}}_i, \quad i = 2 \dots m, \\ \mathbf{s}_j &= \mathbf{L}(\tilde{\mathbf{S}}_j + \mathbf{b}), \quad j = 2 \dots n, \end{aligned} \quad (23)$$

where $\tilde{\mathbf{R}} = \mathbf{L}^T \mathbf{R}$, $\tilde{\mathbf{S}} = \mathbf{L}^{-1} \mathbf{S}$, and hence $\hat{\mathbf{B}} = \tilde{\mathbf{R}}^T \mathbf{L}^{-1} \mathbf{L} \tilde{\mathbf{S}} = \tilde{\mathbf{R}}^T \tilde{\mathbf{S}}$.

Using this parametrization, the equations (19, 20, 21) become

$$d_{11}^2 = \mathbf{b}^T \mathbf{H}^{-1} \mathbf{b}, \quad (24)$$

$$d_{1j}^2 - d_{11}^2 = \tilde{\mathbf{S}}_j^T \mathbf{H}^{-1} \tilde{\mathbf{S}}_j + 2\mathbf{b}^T \mathbf{H}^{-1} \tilde{\mathbf{S}}_j, \quad (25)$$

$$d_{i1}^2 - d_{11}^2 = \tilde{\mathbf{R}}_i^T \mathbf{H} \tilde{\mathbf{R}}_i - 2\mathbf{b}^T \tilde{\mathbf{R}}_i, \quad (26)$$

where we have introduced the symmetric matrix $\mathbf{H} = (\mathbf{L}^T \mathbf{L})^{-1}$.

If enough correspondences are given then we can use the equations of type (26) to estimate \mathbf{H} and \mathbf{b} linearly. Then we can backtrack and calculate $(\mathbf{r}_1, \dots, \mathbf{r}_m, \mathbf{s}_1, \dots, \mathbf{s}_m)$ using equations (23).

If there are both fewer than 10 sound events and fewer than 10 microphones, we can still solve the problem, but we have to use the non-linear equations (24) and (25). This special case can be solved using polynomial equation solving as described in [22].

The initial estimates are then refined by minimizing the non-linear least squares problem,

$$\min_{\mathbf{m}_i, \mathbf{s}_j, o_j} \sum_{ij} (u_{ij} - (\|\mathbf{m}_i - \mathbf{s}_j\|_2 + o_j))^2 \quad (27)$$

using standard techniques (Levenberg-Marquart) in order to obtain the maximal likelihood estimate of the parameters.

7 Find more inliers from time-difference estimation

For find more inliers among the remaining columns in u_{ij} , first we calculate microphone positions \mathbf{r}_i and offsets o_j by trilateration using time differences vector \mathbf{u} and sound position \mathbf{s} . For each time instances in inlier set of \mathbf{u} , trilaterate one point \mathbf{r}_0 using RANSAC followed by bundle adjustment, which is non-linear least square optimization of \mathbf{m}_i with \mathbf{m}_0 as initial estimate.

$$\min_{\mathbf{m}_i} \sum_j (u_{ij} - (\|\mathbf{m}_i - \mathbf{s}_j\|_2)). \quad (28)$$

From I and J which is indeces of inlier set that I represents indeces in channels, 1 to M , and J represents indeces from 1 to number of matches construct vector \mathbf{D} , it takes the values in u which is in inlier set. Then calculate the residuals by $res = \sqrt{[\mathbf{m}(J) - \mathbf{s}(I)]^2} + o(J) - \mathbf{D}$. We set when $res < 0.05$, we count number of inliers in each column if it is ≥ 5 , we set it as inliers.

8 Robust estimation of mirrored microphones

The output for the system so far is typically the microphone positions corresponding to the direct path, i.e. $\mathbf{r}_{1,1}, \dots, \mathbf{r}_{m,1}$ and sound event positions \mathbf{s}_j at a number of different times t_j . Lets return to equation (2),

$$w_{i_1, k_1, i_2, k_2} = \|\mathbf{s} - \mathbf{r}_{i_2, k_2}\| - \|\mathbf{s} - \mathbf{r}_{i_1, k_1}\|. \quad (29)$$

In this section we present a robust method for finding mirrored sound source positions $\mathbf{r}_{i,k}$. From the previous system we have already established some direct path microphone position $\mathbf{r}_{i,1}$ and some sound event positions \mathbf{s} . Lets study the the time-difference measurements from microphone positions to multipath microphone position

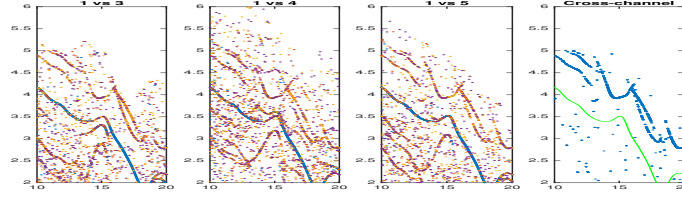


Figure 7: The figure shows GCC-PHAT peaks for three channels corrected for direct path distances according to equation (31) for channel pairs (1 vs 3), (1 vs 4) and (1 vs 5). To the right is shown those data that are consistent over all seven channel pairs (1 vs 2), \dots (1 vs 8). TDOA measurements that are consistent over these channels are candidates for multipath components.

$\mathbf{r}_{i,k}$:

$$\begin{aligned}
 w_{1,1,i,k} &= \|\mathbf{s} - \mathbf{r}_{i,k}\| - \|\mathbf{s} - \mathbf{r}_{1,1}\|, \\
 w_{2,1,i,k} &= \|\mathbf{s} - \mathbf{r}_{i,k}\| - \|\mathbf{s} - \mathbf{r}_{2,1}\|, \\
 &\vdots \\
 w_{m,1,i,k} &= \|\mathbf{s} - \mathbf{r}_{i,k}\| - \|\mathbf{s} - \mathbf{r}_{m,1}\|.
 \end{aligned} \tag{30}$$

By rearrangement

$$\begin{aligned}
 w_{1,1,i,k} + \|\mathbf{s} - \mathbf{r}_{1,1}\| &= \|\mathbf{s} - \mathbf{r}_{i,k}\|, \\
 w_{2,1,i,k} + \|\mathbf{s} - \mathbf{r}_{2,1}\| &= \|\mathbf{s} - \mathbf{r}_{i,k}\|, \\
 &\vdots \\
 w_{m,1,i,k} + \|\mathbf{s} - \mathbf{r}_{m,1}\| &= \|\mathbf{s} - \mathbf{r}_{i,k}\|.
 \end{aligned} \tag{31}$$

we see that we should find the same shape for all of the m different tdoa-images. In Figure 8 we show the top 4 peaks of the GCC-PHAT signal for each time instant corrected for direct path distance according to equation (31). Notice that some curves are identical in the peaks. The cross channel consistent data is shown to the right.

Using the cross channel consistent data (Figure 8) we get candidates for multipath component distances $\|\mathbf{s} - \mathbf{r}_{i,k}\|$. Again since we already have an estimate on sound source paths, the problem has only three degrees of freedom in, i.e. the components of $\mathbf{r}_{i,k}$. Again we use a hypothesize and test paradigm.

1. Choose three random points among the cross channel consistent data.
2. Find the two candidate positions for the multi-path position of the microphone.
3. For each candidate position \mathbf{r} calculate $\|\mathbf{s} - \mathbf{r}\|$ and check for inliers.
4. Repeat steps 1-3.
5. Choose the point \mathbf{r} with the most inliers.

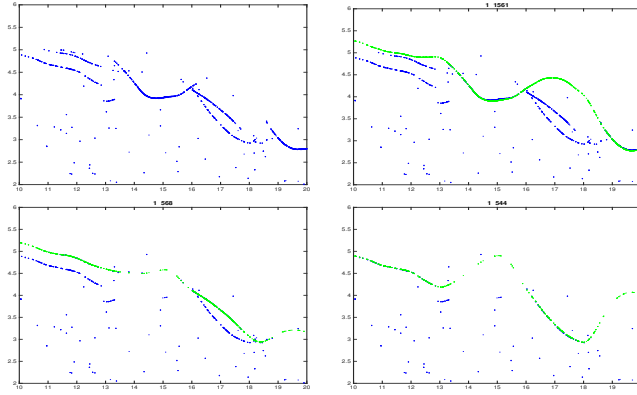


Figure 8: The figure shows the candidates for multipath components (top-left). After running RANSAC once we obtain a candidate for a multi-path microphone position consistent with 1561 inliers (top-right). After running RANSAC once more on the remaining data we obtain another candidate for a multi-path microphone position consistent with 564 inliers (bottom-left). After running RANSAC once more on the remaining data we obtain another candidate for a multi-path microphone position consistent with 544 inliers (bottom-right).

6. Perform local non-linear optimization to minimize sum of squared reprojection errors among the inliers.

After finding one mirrored position remove the inlier data and repeat the process.

This process is then repeated for the other channels so as to find multi-path mirrored microphone positions for all microphones.

9 Experimental Validation

We have made several experiments to verify the different components of the system as well as the robustness and performance of the whole system.

We first study the numerical stability of the minimal solvers (Sections 5-6). Then we study the performance of the tracking algorithms of Section 4. Then follows evaluations of the whole system for a couple of different scenarios.

9.1 Synthetic Data

In this section, we study the numerical behaviors of the TDOA solvers on synthetic data. We simulate the positions of microphones and sounds as 3D points with independent Gaussian distribution of zero mean and identity covariance matrix. As for the offsets, we choose them from independent Gaussian distribution with zero mean and standard deviation 10. We study the effects of zero-mean Gaussian noise on the solvers, where we vary the standard deviation of the Gaussian noise added to the TDOA measurements. When solving TOA problem, we have used the scheme discussed in Section

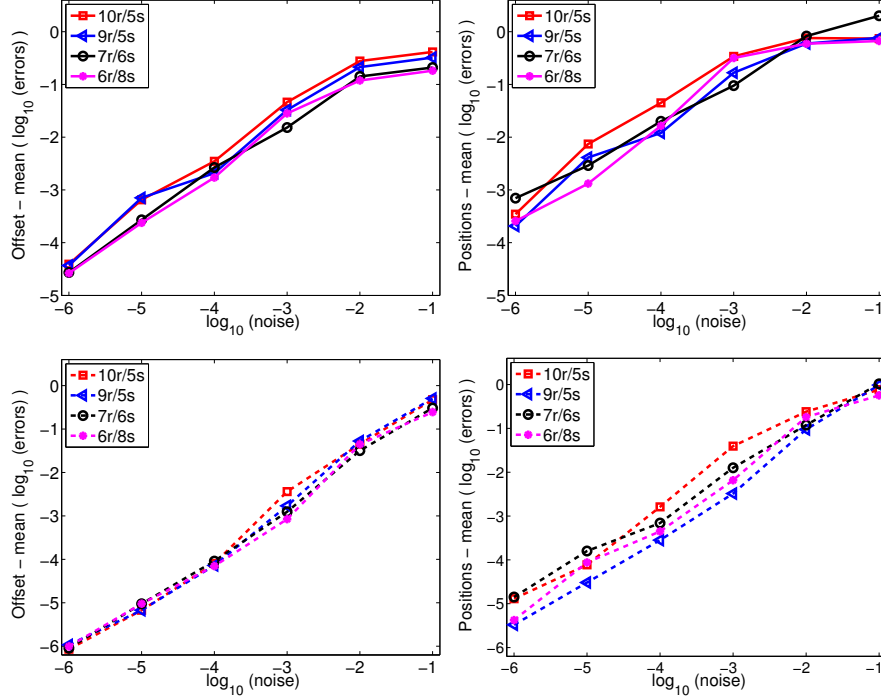


Figure 9: Synthetic experiments for TDOA solvers on 3D under Gaussian noise. The errors of estimated time offsets (left) and reconstructed positions of microphones and sounds (right) are shown. (Top) Performance of different solvers (10r/5s [24], 9r/5s, 7r/6s and 6r/8s) with their corresponding minimal settings for solving offsets; (Bottom) with 20 receivers and 20 microphones.

6 for over-determined cases. To compare the reconstructed positions of microphones and sounds with the true positions, we rotate and translate the coordinate system accordingly. We can see that from Fig. 9, our proposed solvers 9r/5s, 7r/6s and 6r/8s give numerically similar results as the 10r/5s case in [24] for both minimal settings and over-determined cases. In Fig.10, random initialization of the time offsets resulted in poor convergence in the non-linear optimization, while our method provides with a much better starting point. On the other hand, we have also compared our solver with the iterative method proposed in [36] for estimating the offsets. The method in [36] converges very slowly (5sec. - 1min. on a standard laptop, especially for (near) minimal settings) and tends to converge to the wrong local minima. While our solvers perform consistently well for all cases, they are also much faster (approximately 0.5s for the unoptimized codes).

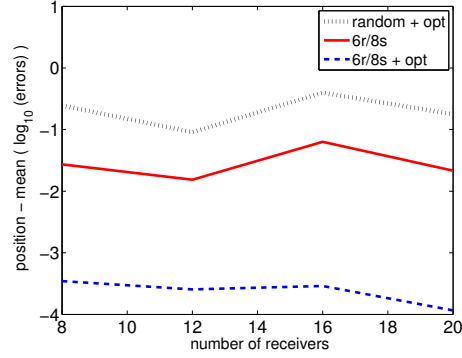


Figure 10: Initialization with random offsets and our 6r/8s solver with varying number receivers (8 to 20) and 8 transmitters (noise level 10^{-4}) for non-linear optimization.

9.2 Claps

For this experiment we used 8 Shure SV100 microphones recorded at 44.1kHz in an office. They are approximately 0.3-1.5 meters away from each other and are placed so that they span 3D. We connected them to an audio interface (M-Audio Fast Track Ultra 8R), which is connected to a computer. We generated sounds by moving around in the room and clapping approximately 1-2 meters from the microphones. We collected 5 independent recordings of approximately 20s. Each recording contained roughly 30 claps (sound events).

To obtain TDOA measurements, we coarsely matched sounds of the claps to sound flanks as described in Section 4.A. For the experiment we used only those claps that were detected in all 8 channels. We ran both the 7r/6s and 6r/8s solvers to determine the offsets followed by an alternating optimization that refines the offset estimation. After solving the unknown transformation and translation, we recover an initial euclidean reconstruction for the locations of microphones and claps. Finally we refine the reconstruction with non-linear optimization. The result of one of these 5 reconstructions are shown in Figure 11 (middle). The reconstructed microphone positions from these 5 independent multi-channel recordings were put in a common coordinate system and compared to each other. The average distance from each microphone to the its corresponding mean position (estimated from corresponding reconstruction of the 5 recordings) is 2.60 cm. It is important to point out that without proper initialization using our methods, the solutions we get converge poorly (with large reconstruction errors). Previous solvers do not work here due to either insufficient number of receivers (10 receivers needed in [24]) or violating the assumption that one of the microphones collocates with one of the claps [41].

As an additional evaluation, we have also reconstructed the locations of the microphones based on computer vision techniques. We took 11 images of the experimental setup. Figure 11 (left) shows one of the 11 images used. We manually detected the 8 microphone center positions in these 11 images and used standard structure from motion algorithms to estimate the positions of the 8 microphones. The resulting recon-

struction is also compared to that of the five structure from sound reconstructions. The comparison is shown in Figure 11 (right). We can see the TDOA-based reconstructions are consistent with the vision-based reconstruction.

9.3 Continuous songs in anechoic chambers

We have made several experiments using a continuously moving sound source in anechoic chambers. These produced GCC-PHAT data that were relatively easy for our system to track and match over time. Even in the case of two sound sources moving simultaneously and continuously in the chamber produced relatively good results.

We illustrate some of the steps of the automatic system with one of the experiments. In this case we have two moving sound sources in 3D. We assume $c = 343 \text{ m/s}$ in room temperature. The matching algorithm produces 129 matching vectors. There are 83 missing data among these $129 \times 8 = 1032$ time difference measurements. The RANSAC algorithm finds an inlier set of 75 (out of the 129) matching vectors.

These 75 inlier matching vectors are then used to estimate the 3D positions for the senders and receivers. A histogram of the residuals $u_{ij} - (||\mathbf{m}_i - \mathbf{s}_j||_2 + o_j)$ is shown in Figure 13. The errors are in the order of a few millimeters. The final 3D reconstruction of the microphones and of the sound source paths for one of the experiments are shown in Figures 12. In this experiment we have microphones in two planes (four in each). The moving sound source starts outside the convex hull of the microphones, then moves inside the microphone cluster and then out again.

After refinement step, we get 129 matches.

To validate the method we have used several independent recordings. These have different sound source positions, but identical microphone setup. The error between the three reconstructions and the mean has a standard deviation of about 1 cm.

9.4 Matching in reverberant surroundings

Several experiments were carried out. During these experiments we used eight T-bone MM-1 microphones, connected to a M-Audio Fast Track Ultra 8R sound card and recorded with the program Audacity on a MacBook Pro with 3 GHz Intel Core i7 processor. Recordings were made in different rooms (seminar rooms, lunch rooms etc). The experimental environments were chosen as to have a few planar surfaces (floor, walls, ceilings) with relatively low sound absorption coefficient as well as a few other objects such as chairs, tables, flower pots etc. Several experiments were made with a single moving sound source, i.e. a Roxcore Portasound loudspeaker connected to a mobile phone. In the experiment we used different songs.

For the experiment we generated ground truth matching difference vectors for the direct path. After each of the steps of the matching system, as described in Section 4, we computed the detections obtained by the automatic system with the ground truth and calculated the number of inlier detections and outlier detections. These are shown in Table 2. Notice that the number of inliers are kept at a high level, while the number of outliers decrease. The output of the system is essentially outlier free.

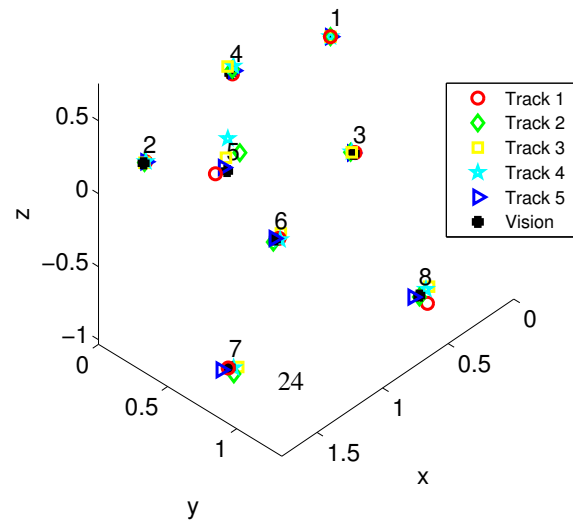
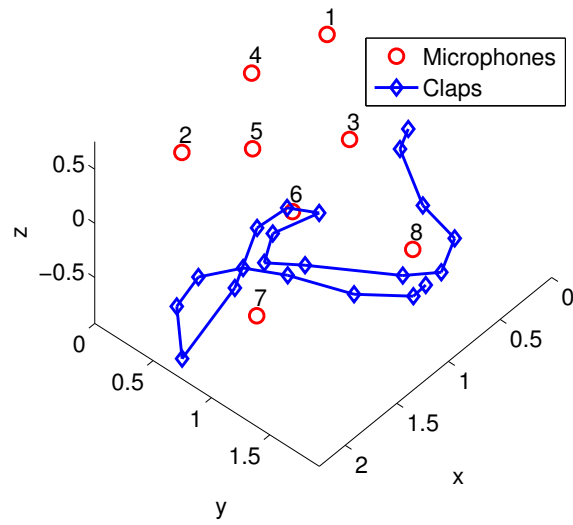


Figure 11: Results on TDOA with microphones and sounds. Top: View of the experimental setup. Middle : Reconstruction of 8 microphones (in red - 'o') and 21

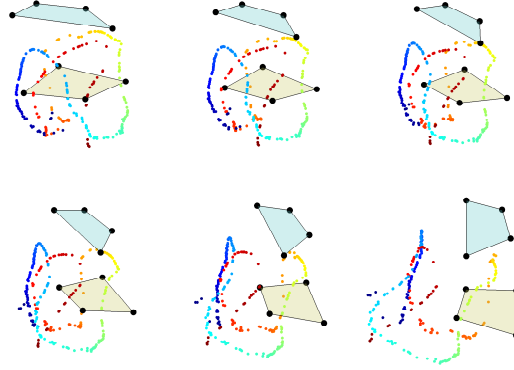


Figure 12: Three-dimensional reconstruction of the microphone setup (black dots) as well as three-dimensional reconstruction of moving sound source positions (coloured dots). Note that the microphones in the experiment setup are located in two planes (also indicated in the figure).

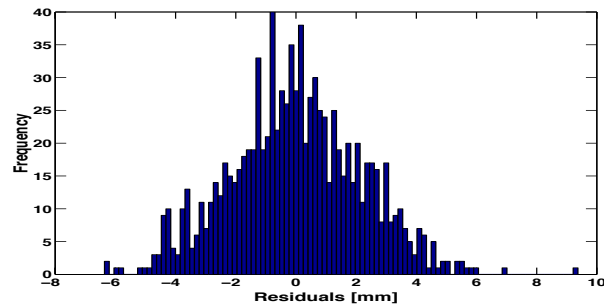


Figure 13: Histogram of the residuals between the measured data u_{ij} and the fit $\|\mathbf{m}_i - \mathbf{s}_j\|_2 + o_j$.

Step	Inliers	Outliers	Corresponding figure
Step 1	1260	4639	
Step 2	1228	633	Figure 4
Step 3	1027	30	Figure 5
Step 4	1199	93	
Step 5	1199	7	
Step 6	1475	17	Figure 6
Step 7	1275	0	

Table 2: The number of outliers and inliers after each step of the algorithm.

9.5 System evaluation

The detections of the system is then used as input to the estimation of the parameters \mathbf{r}_i , \mathbf{s}_j and \mathbf{o}_j as described in Sections 5-6. For visualization purposes we have calculated the camera matrix P for the image in relation to the world coordinate system obtained by the audio reconstruction. This makes it possible to visualize the estimated motion path of the sound source in the video. We also made several independent recordings with the microphones in the same positions. The reconstructed microphone positions could then be compared with each other, thus estimating the reconstruction errors. The estimated standard deviation in microphone coordinates was 3.7 millimeters.

Mirrored microphones

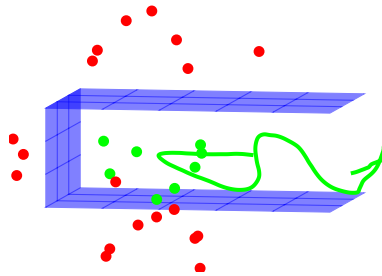


Figure 14: The figure illustrates the estimated microphone positions (green dots), estimated mirrored microphone positions (red dots) and sound source motion (green curve). The estimated reflective planes are also shown in the figure. These three planes correspond to the floor, the ceiling and the walls as seen in Figure 1

10 Conclusion

In this paper, we have developed a novel automatic microphone self-localization using ambient sound. The system does not put any constraints on the motion of the sound sources in relation to the microphone array setup. The system is based on a first finding several time-difference matching vectors for the recording. These are then used as input to robust geometric algorithms based on minimal solvers and RANSAC to provide initial estimates of the unknown parameters, i.e. the offsets and the 3D positions of the sound sources and the receivers. For this purpose we have developed efficient algorithms that solve for offsets using minimal data. There are three such minimal problems assuming that microphones and sound sources span 3D. There are two such minimal problems assuming that either the microphones or the sound sources are restricted to a plane. For the case of microphones or sound sources restricted to a line there is one such minimal problem. The robust estimation algorithm produces an initial estimate of the geometry parameters and a set of inlier measurements. These estimates of the geometry parameters are then improved by non-linear optimization using the inlier data to obtain the maximum likelihood estimate of the parameters. We have also used sub-sample refinement of the time-difference-of-arrival measurements and shown that they improve the precision of the system. The resulting microphone accuracy is in the order of millimeters. The components of the system as well as the system as a whole has been tested on both synthetic and real data with promising results.

Although we have solved the minimal problems for offset estimation using a rank constraint, it is still an open problem to study and solve the minimal problems for solving the TDOA problem using all constraints. Other possible avenues for future research would be to improve on the overall system, making it even more robust. Here different strategies for tracking could be tried out on real and synthetic data. The code for the minimal solvers as well as the system is published on github, <http://github.com/kalleastrom/StructureFromSound>. We also provide on our homepage a set of example recordings that can be used with the system, <http://vision.maths.lth.se/sfsdb/>. We hope that this can be used to improve on the system.

References

- [1] Yubin Kuang and Kalle Åström, “Stratified sensor network self-calibration from tdoa measurements,” in *21st European Signal Processing Conference 2013*, 2013.
- [2] Simayijiang Zhayida, Fredrik Andersson, Yubin Kuang, and Kalle Åström, “An automatic system for microphone self-localization using ambient sound,” in *22st European Signal Processing Conference*, 2014.
- [3] Axel Plinge, Florian Jacob, Reinhold Haeb-Umbach, and Gernot A Fink, “Acoustic microphone geometry calibration: An overview and experimental evaluation of state-of-the-art algorithms,” *IEEE Signal Processing Magazine*, vol. 33, no. 4, pp. 14–29, 2016.

- [4] M.S. Brandstein, J.E. Adcock, and H.F. Silverman, "A closed-form location estimator for use with room environment microphone arrays," *Speech and Audio Processing, IEEE Transactions on*, vol. 5, no. 1, pp. 45–50, Jan. 1997.
- [5] A. Cirillo, R. Parisi, and A. Uncini, "Sound mapping in reverberant rooms by a robust direct method," in *Acoustics, Speech and Signal Processing, IEEE International Conference on*, April 2008, pp. 285–288.
- [6] M. Cobos, A. Marti, and J.J. Lopez, "A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling," *Signal Processing Letters, IEEE*, vol. 18, no. 1, pp. 71–74, Jan. 2011.
- [7] Hoang Do, H.F. Silverman, and Ying Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction(SRC) on a large-aperture microphone array," in *ICASSP 2007*, April 2007, vol. 1, pp. 121–124.
- [8] C. Knapp and G.Clifford Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.
- [9] Simon Burgess, Yubin Kuang, and Kalle Åström, "{TOA} sensor network self-calibration for receiver and transmitter spaces with difference in dimension," *Signal Processing*, vol. 107, pp. 33–42, 2015, Special Issue on ad hoc microphone arrays and wireless acoustic sensor networksSpecial Issue on Fractional Signal Processing and Applications.
- [10] M. Batstone, K. abd Oskarsson and K. Åström, "Robust time-of-arrival self calibration and indoor localization using wi-fi round-trip time measurements," in *proc. of International Conference on Communication*, 2016.
- [11] S. T. Birchfield and A. Subramanya, "Microphone array position calibration by basis-point classical multidimensional scaling," *IEEE transactions on Speech and Audio Processing*, vol. 13, no. 5, 2005.
- [12] D. Niculescu and B. Nath, "Ad hoc positioning system (aps)," in *GLOBECOM-01*, 2001.
- [13] V. C. Raykar, I. V. Kozintsev, and R. Lienhart, "Position calibration of microphones and loudspeakers in distributed computing platforms," *IEEE transactions on Speech and Audio Processing*, vol. 13, no. 1, 2005.
- [14] M. Crocco, A. Del Bue, M. Bustreo, and V. Murino, "A closed form solution to the microphone position self-calibration problem," in *ICASSP*, March 2012.
- [15] J. C. Chen, R. E. Hudson, and K. Yao, "Maximum likelihood source localization and unknown sensor location estimation for wideband signals in the near-field," *IEEE transactions on Signal Processing*, vol. 50, 2002.

- [16] P. Pertila, M.S. Hamalainen, and M. Mieskolainen, "Passive temporal offset estimation of multichannel recordings of an ad-hoc microphone array," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 11, pp. 2393–2402, Nov. 2013.
- [17] R. Biswas and S. Thrun, "A passive approach to sensor network localization," in *IROS 2004*, 2004.
- [18] J. Wendeborg, F. Hoflinger, C. Schindelhauer, and L. Reindl, "Anchor-free TDOA self-localization," in *Indoor Positioning and Indoor Navigation (IPIN), 2011 International Conference on*, Sept. 2011, pp. 1–10.
- [19] P. Biswas, T.C. Lian, T.C. Wang, and Y. Ye, "Semidefinite programming based algorithms for sensor network localization," *ACM Transactions on Sensor Networks (TOSN)*, vol. 2, no. 2, pp. 188–220, 2006.
- [20] Y. Kuang, E. Ask, S. Burgess, and K. Åström, "Understanding toa and tdoa network calibration using far field approximation as initial estimate," in *ICPRAM*, 2012.
- [21] H. Stewénius, *Gröbner Basis Methods for Minimal Problems in Computer Vision*, Ph.D. thesis, Lund University, April 2005.
- [22] Y. Kuang, S. Burgess, A. Torstensson, and K. Åström, "A complete characterization and solution to the microphone position self-calibration problem," in *ICASSP*, 2013.
- [23] S. Thrun, "Affine structure from sound," in *Proceedings of Conference on Neural Information Processing Systems*, 2005.
- [24] M. Pollefeys and D. Nister, "Direct computation of sound and microphone locations from time-difference-of-arrival data," in *Proc. of ICASSP*, 2008.
- [25] SD Valente, F Antonacci, M Tagliasacchi, A Sarti, and S Tubaro, "Self-calibration of two microphone arrays from volumetric acoustic maps in non-reverberant rooms," in *Communications, Control and Signal Processing (ISCCSP), 2010 4th International Symposium on*. IEEE, 2010, pp. 1–4.
- [26] Joerg Schmalenstroeer, Florian Jacob, Reinhold Haeb-Umbach, Marius Hennecke, and Gernot A Fink, "Unsupervised geometry calibration of acoustic sensor networks using source correspondences," in *Interspeech 2011*, 2011.
- [27] O. D. Faugeras, L. Quan, and P. Sturm, "Self-calibration of a 1d projective camera and its application to the self-calibration of a 2d projective camera," in *Proc. 5th European Conf. on Computer Vision, Freiburg, Germany*. 1998, pp. 36–52, Springer-Verlag.
- [28] K. Åström and M. Oskarsson, "Solutions and ambiguities of the structure and motion problem for 1d retinal vision," *Journal of Mathematical Imaging and Vision*, vol. 12, no. 2, pp. 121–135, 2000.

- [29] K. Åström and F. Kahl, “Ambiguous configurations for the 1d structure and motion problem,” *Journal of Mathematical Imaging and Vision*, vol. 18, no. 2, pp. 191–203, 2003.
- [30] Iain McCowan, Mike Lincoln, and Ivan Himawan, “Microphone array shape calibration in diffuse noise fields,” *IEEE Transactions on Audio, Speech & Language Processing*, vol. 16, no. 3, pp. 666–670, 2008.
- [31] G. Young and A.S. Householder, “Discussion of a set of points in terms of their mutual distances,” *Psychometrika*, vol. 3, no. 1, pp. 19–22, 1938.
- [32] Marius Hennecke, Thomas Plotz, Gernot A Fink, Jorg Schmalenstroer, and Reinhold Hab-Umbach, “A hierarchical approach to unsupervised shape calibration of microphone array networks,” in *2009 IEEE/SP 15th Workshop on Statistical Signal Processing*. IEEE, 2009, pp. 257–260.
- [33] K. Åström and A. Heyden, “Stochastic analysis of scale-space smoothing,” *Advances in Applied Probability*, vol. 30, no. 1, 1999.
- [34] Simon Burgess, Yubin Kuang, Johannes Wendeberg, Kalle Åström, and Christian Schindelhauer, “Minimal solvers for unsynchronized tdoa sensor network calibration,” in *International Symposium on Algorithms and Experiments for Sensor Systems, Wireless Networks and Distributed Robotics*. Springer, 2013, pp. 95–110.
- [35] Shuanglong Liu, Chun Zhang, and Yu Huang, “Research on acoustic source localization using time difference of arrival measurements,” in *Measurement, Information and Control (MIC), 2012 International Conference on*. IEEE, 2012, vol. 1, pp. 220–224.
- [36] F. Jiang, Y. Kuang, and K. Åström, “Time delay estimation for tdoa self-calibration using truncated nuclear norm,” in *Proc. of ICASSP*, 2013.
- [37] Daniel R. Grayson and Michael E. Stillman, “Macaulay2,” <http://www.math.uiuc.edu/Macaulay2/>.
- [38] M. Byröd, K. Josephson, and K. Åström, “Fast and stable polynomial equation solving and its application to computer vision,” *IJCV*, 2009.
- [39] M. A. Fischler and R. C. Bolles, “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography,” *Communications of the ACM*, vol. 24, no. 6, pp. 381–95, 1981.
- [40] N. D. Gaubitch, W. B. Kleijn, and R. Heusdens, “Auto-localization in ad-hoc microphone arrays,” in *ICASSP*, 2013.
- [41] M. Crocco, A. Del Bue, and V. Murino, “A bilinear approach to the position self-calibration of multiple sensors,” *Trans. Sig. Proc.*, vol. 60, no. 2, pp. 660–673, feb 2012.